

---

# Fast Convergence Rate of Multiple Kernel Learning with Elastic-net Regularization

---

**Taiji Suzuki, Ryota Tomioka**

Department of Mathematical Informatics,  
The University of Tokyo,  
7-3-1 Hongo, Bunkyo-ku, Tokyo  
t-suzuki@mist.i.u-tokyo.ac.jp,  
tomioka@mist.i.u-tokyo.ac.jp

**Masashi Sugiyama**

Department of Computer Science,  
Tokyo Institute of Technology,  
2-12-1 O-okayama, Meguro-ku, Tokyo  
sugi@cs.titech.ac.jp

## Abstract

We investigate the learning rate of multiple kernel learning (MKL) with elastic-net regularization, which consists of an  $\ell_1$ -regularizer for inducing the sparsity and an  $\ell_2$ -regularizer for controlling the smoothness. We focus on a sparse setting where the total number of kernels is large but the number of non-zero components of the ground truth is relatively small, and prove that elastic-net MKL achieves the minimax learning rate on the  $\ell_2$ -mixed-norm ball. Our bound is sharper than the convergence rates ever shown, and has a property that the smoother the truth is, the faster the convergence rate is.

## 1 Introduction

Learning with kernels such as support vector machines has been demonstrated to be a promising approach, given that kernels were chosen appropriately (Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004). So far, various strategies have been employed for choosing appropriate kernels, ranging from simple cross-validation (Chapelle et al., 2002) to more sophisticated ‘kernel learning’ approaches (Ong et al., 2005, Argyriou et al., 2006, Bach, 2009, Cortes et al., 2009a, Varma and Babu, 2009).

*Multiple kernel learning* (MKL) is one of the systematic approaches to learning kernels, which tries to find the optimal linear combination of prefixed base-kernels by convex optimization (Lanckriet et al., 2004). The seminal paper by Bach et al. (2004) showed that this linear-combination MKL formulation can be interpreted as  $\ell_1$ -mixed-norm regularization (i.e., the sum of the norms of the base kernels). Based on this interpretation, several variations of MKL were proposed, and promising performance was achieved by ‘intermediate’ regularization strategies between the sparse ( $\ell_1$ ) and dense ( $\ell_2$ ) regularizers, e.g., a mixture of  $\ell_1$ -mixed-norm and  $\ell_2$ -mixed-norm called the *elastic-net regularization* (Shawe-Taylor, 2008, Tomioka and Suzuki, 2009) and  $\ell_p$ -mixed-norm regularization with  $1 < p < 2$  (Micchelli and Pontil, 2005, Kloft et al., 2009).

Together with the active development of practical MKL optimization algorithms, theoretical analysis of MKL has also been extensively conducted. For  $\ell_1$ -mixed-norm MKL, Koltchinskii and Yuan (2008) established the learning rate  $d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} + d \log(M)/n$  under rather restrictive conditions, where  $n$  is the number of samples,  $d$  is the number of non-zero components of the ground truth,  $M$  is the number of kernels, and  $s$  ( $0 < s < 1$ ) is a constant representing the complexity of the reproducing kernel Hilbert spaces (RKHSs). Their conditions include a smoothness assumption of the ground truth ( $q = 1$  in our terminology (Assumption 2)). For elastic-net regularization, Meier et al. (2009) gave a near optimal convergence rate  $d(n/\log(M))^{-\frac{1}{1+s}}$ . Recently, Koltchinskii and Yuan (2010) showed that MKL with a variant of  $\ell_1$ -mixed-norm regularization achieves the minimax optimal convergence rate, which successfully got a sharper dependency with respect to  $\log(M)$  than the bound of Meier et al. (2009) and established the bound  $dn^{-\frac{1}{1+s}} + d \log(M)/n$ . Another line of research considers the cases where the ground truth is not sparse, and bounds the Rademacher complexity of a candidate kernel class by a pseudo-dimension of the kernel class (Srebro and Ben-David, 2006, Ying and Campbell, 2009, Cortes et al., 2009b, Kloft et al., 2010).

In this paper, we focus on the sparse setting (i.e., the total number of kernels is large, but the number of non-zero components of the ground truth is relatively small), and derive a sharp learning

rate for elastic-net MKL. Our new learning rate,

$$d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d \log(M)}{n},$$

is faster than all the existing bounds, where  $R_{2,g^*}$  is a kind of the  $\ell_2$ -mixed-norm of the truth and  $q$  ( $0 \leq q \leq 1$ ) is a constant depending on the smoothness of the ground truth.

Our contributions are summarized as follows.

- The sharpest existing bound given by Koltchinskii and Yuan (2010) achieves the minimax rate on the  $\ell_\infty$ -mixed-norm ball (Raskutti et al., 2009, 2010). Our work follows this line and show that the learning rate for elastic-net MKL further achieves the minimax rate on the  $\ell_2$ -mixed-norm ball, which is faster than that on the  $\ell_\infty$ -mixed-norm ball. This result implies that the bound by Koltchinskii and Yuan (2010) is tight only when the ground truth is evenly spread in the non-zero components.
- We included the *smoothness*  $q$  of the ground truth into our learning rate, where the ground truth is said to be smooth if it is represented as a convolution of a certain function and an integral kernel (see Assumption 2). Intuitively for larger  $q$ , the truth is smoother. We show that, the smoother the truth is, the faster the convergence rate is. That is, the resultant convergence rate becomes as if the complexity of RKHSs was  $\frac{s}{1+q}$  instead of the true complexity  $s$ . Meier et al. (2009), Koltchinskii and Yuan (2010) assumed  $q = 0$  and Koltchinskii and Yuan (2008) considered a situation of  $q = 1$ . Our analysis covers those situations.

## 2 Preliminaries

In this section, we formulate elastic-net MKL, and summarize mathematical tools that are needed for theoretical analysis.

### 2.1 Formulation

Suppose we are given  $n$  samples  $(x_i, y_i)_{i=1}^n$  where  $x_i$  belongs to an input space  $\mathcal{X}$  and  $y_i \in \mathbb{R}$ . We denote the marginal distribution of  $X$  by  $\Pi$ . We consider a MKL regression problem in which the unknown target function is represented as a form of  $f(x) = \sum_{m=1}^M f_m(x)$  where each  $f_m$  belongs to a different RKHS  $\mathcal{H}_m$  ( $m = 1, \dots, M$ ) with kernel  $k_m$  over  $\mathcal{X} \times \mathcal{X}$ .

The elastic-net MKL we consider in this paper is the version considered in Meier et al. (2009):

$$\hat{f} = \arg \min_{\substack{f_m \in \mathcal{H}_m \\ (m=1, \dots, M)}} \frac{1}{n} \sum_{i=1}^N \left( y_i - \sum_{m=1}^M f_m(x_i) \right)^2 + \sum_{m=1}^M \lambda_1^{(n)} \sqrt{\|f_m\|_n^2 + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2, \quad (1)$$

where  $\|f_m\|_n := \sqrt{\frac{1}{n} \sum_{i=1}^n f_m(x_i)^2}$  and  $\|f_m\|_{\mathcal{H}_m}$  is the RKHS norm of  $f_m$  in  $\mathcal{H}_m$ . The regularizer is the mixture of  $\ell_1$ -term  $\sum_m \sqrt{\|f_m\|_n^2 + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}^2}$  and  $\ell_2$ -term  $\sum_m \|f_m\|_{\mathcal{H}_m}^2$ . In that sense, we say that the regularizer is of the elastic-net type<sup>1</sup> (Zou and Hastie, 2005). Here the  $\ell_1$  term is a mixture of the empirical  $L_2$  norm  $\|f_m\|_n$  and the RKHS norm  $\|f_m\|_{\mathcal{H}_m}$ . Koltchinskii and Yuan (2010) also considered  $\ell_1$  regularization that is a mixture of these quantities:  $\sum_m \lambda_1^{(n)} \|f_m\|_n + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}$ .

By the representer theorem (Kimeldorf and Wahba, 1971), the solution  $\hat{f}$  can be expressed as a linear combination of  $nM$  kernels:  $\exists \alpha_{m,i} \in \mathbb{R}$ ,  $\hat{f}_m(x) = \sum_{i=1}^n \alpha_{m,i} k_m(x, x_i)$ . Thus, using the Gram matrix  $\mathbf{K}_m = (k_m(x_i, x_j))_{i,j}$ , the regularizer in (1) is expressed as

$$\sum_{m=1}^M \lambda_1^{(n)} \sqrt{\boldsymbol{\alpha}_m^\top \left( \frac{\mathbf{K}_m \mathbf{K}_m}{n} + \lambda_2^{(n)} \mathbf{K}_m \right) \boldsymbol{\alpha}_m} + \lambda_3^{(n)} \sum_{m=1}^M \boldsymbol{\alpha}_m^\top \mathbf{K}_m \boldsymbol{\alpha}_m,$$

where  $\boldsymbol{\alpha}_m = (\alpha_{m,i})_{i=1}^n \in \mathbb{R}^n$ . Thus, we can solve the problem by a SOCP (second-order cone programming) solver as in Bach et al. (2004), or the coordinate descent algorithms (Meier et al., 2008).

<sup>1</sup> There is another version of MKL with elastic-net regularization considered in Shawe-Taylor (2008) and Tomioka and Suzuki (2009), that is,  $\lambda_1^{(n)} \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m} + \lambda_2^{(n)} \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2$  (i.e., there is no  $\|f_m\|_n$  term in the regularizer). However, we focus on the former one because the later one is too loose to properly bound the irrelevant components of the estimated function.

## 2.2 Notations and Assumptions

Here, we present several assumptions used in our theoretical analysis and prepare notations.

Let  $\mathcal{H} = \mathcal{H}_1 \oplus \dots \oplus \mathcal{H}_M$ . We denote by  $f^* \in \mathcal{H}$  the ground truth satisfying the following assumption.

### Assumption 1 (Basic Assumptions)

(A1-1) *There exists  $f^* = (f_1^*, \dots, f_M^*) \in \mathcal{H}$  such that  $E[Y|X] = \sum_{m=1}^M f_m^*(X)$ , and the noise  $\epsilon := Y - f^*(X)$  is bounded as  $|\epsilon| \leq L$ .*

(A1-2) *For each  $m = 1, \dots, M$ ,  $\mathcal{H}_m$  is separable and  $\sup_{X \in \mathcal{X}} |k_m(X, X)| \leq 1$ .*

The first assumption in (A1-1) ensures the model  $\mathcal{H}$  is correctly specified, and the technical assumption  $|\epsilon| < L$  allows  $\epsilon f$  to be Lipschitz continuous with respect to  $f$ . These assumptions are not essential and can be relaxed to misspecified models and unbounded noise such as Gaussian noise (Raskutti et al., 2010). However, for the sake of simplicity, we assume these conditions.

It is known that the assumption (A1-2) gives the following relation:

$$\|f_m\|_\infty \leq \sup_x \langle k_m(x, \cdot), f_m \rangle_{\mathcal{H}_m} \leq \sup_x \|k_m(x, \cdot)\|_{\mathcal{H}_m} \|f_m\|_{\mathcal{H}_m} \leq \sup_x \sqrt{k_m(x, x)} \|f_m\|_{\mathcal{H}_m} \leq \|f_m\|_{\mathcal{H}_m}.$$

Later, we will also assume a stronger (but practical) condition on the sup-norm in Assumption 5.

We define an operator  $T_m : \mathcal{H}_m \rightarrow \mathcal{H}_m$  as

$$\langle f_m, T_m g_m \rangle_{\mathcal{H}_m} := E[f_m(X)g_m(X)],$$

where  $f_m, g_m \in \mathcal{H}_m$ . Due to Mercer's theorem, there are an orthonormal system  $\{\phi_{k,m}\}_{k,m}$  in  $L_2(\Pi)$  and the spectrum  $\{\mu_{k,m}\}_{k,m}$  such that  $k_m$  has the following spectral representation:

$$k_m(x, x') = \sum_{k=1}^{\infty} \mu_{k,m} \phi_{k,m}(x) \phi_{k,m}(x'). \quad (2)$$

By this spectral representation, the inner-product of RKHS can be expressed as  $\langle f_m, g_m \rangle_{\mathcal{H}_m} = \sum_{k=1}^{\infty} \mu_{k,m}^{-1} \langle f_m, \phi_{k,m} \rangle_{L_2(\Pi)} \langle \phi_{k,m}, g_m \rangle_{L_2(\Pi)}$ .

**Assumption 2 (Convolution Assumption)** *There exist a real number  $0 \leq q \leq 1$  and  $g_m^* \in \mathcal{H}_m$  such that*

$$(A2) \quad f_m^*(x) = \int_{\mathcal{X}} k_m^{(q/2)}(x, x') g_m^*(x') d\Pi(x') \quad (\forall m = 1, \dots, M),$$

where  $k_m^{(q/2)}(x, x') = \sum_{k=1}^{\infty} \mu_{k,m}^{q/2} \phi_{k,m}(x) \phi_{k,m}(x')$ . This is equivalent to the following operator representation:

$$f_m^* = T_m^{\frac{q}{2}} g_m^*.$$

The constant  $q$  controls the smoothness of the truth  $f_m^*$  because  $f_m^*$  is a convolution of the integral kernel  $k_m^{(q/2)}$  and  $g_m^*$ , and high frequency components are depressed as  $q$  becomes large. Therefore, as  $q$  becomes large,  $f^*$  becomes “smooth”. The assumption (A2) was considered in Caponnetto and de Vito (2007) to analyze the convergence rate of least-squares estimators in a single kernel setting. In MKL settings, Koltchinskii and Yuan (2008) showed a fast learning rate of MKL, and Bach (2008) employed the assumption for  $q = 1$  to show the consistency of MKL. Proposition 9 of Bach (2008) gave a sufficient condition to fulfill (A2) with  $q = 1$  for translation invariant kernels  $k_m(x, x') = h_m(x - x')$ . Meier et al. (2009) considered a situation with  $q = 0$  on Sobolev space; the analysis of Koltchinskii and Yuan (2010) also corresponds to  $q = 0$ . Note that (A2) with  $q = 0$  imposes nothing on the smoothness about the truth, and our analysis also covers this case.

We will show in Appendix A that as  $q$  increases, the space of the functions that satisfy (A2) becomes “simple”. Thus, it might be natural to consider that, under the Convolution Assumption (A2), the learning rate becomes faster as  $q$  increases. Although this conjecture is actually true, it is not obvious because the Convolution Assumption only restricts the ground truth, but not the search space.

Next we introduce a parameter representing the complexity of RKHSs.

**Assumption 3 (Spectral Assumption)** *There exist  $0 < s < 1$  and  $c$  such that*

$$(A3) \quad \mu_{k,m} \leq ck^{-\frac{1}{s}}, \quad (1 \leq k, 1 \leq m \leq M),$$

where  $\{\mu_{k,m}\}_k$  is the spectrum of the kernel  $k_m$  (see Eq.(2)).

It was shown that the spectral assumption (A3) is equivalent to the classical covering number assumption<sup>2</sup> (Steinwart et al., 2009). If the spectral assumption (A3) holds, there exists a constant  $C$  that depends only on  $s$  and  $c$  such that

$$\mathcal{N}(\varepsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi)) \leq C\varepsilon^{-2s}, \quad (3)$$

and the converse is also true (see Theorem 15 of Steinwart et al. (2009) and Steinwart (2008) for details). Therefore, if  $s$  is large, at least one of the RKHSs is “complex”, and if  $s$  is small, all the RKHSs are “simple”. A more detailed characterization of the covering number in terms of the spectrum is provided in Appendix A. The covering number of the space of functions that satisfy the Convolution Assumption (A2) is also provided there.

We denote by  $I_0$  the indices of truly active kernels, i.e.,

$$I_0 := \{m \mid \|f_m^*\|_{\mathcal{H}_m} > 0\}.$$

For  $f = \sum_{m=1}^M f_m \in \mathcal{H}$  and a subset of indices  $I \subseteq \{1, \dots, M\}$ , we define  $\mathcal{H}_I = \oplus_{m \in I} \mathcal{H}_m$  and denote by  $f_I \in \mathcal{H}_I$  the restriction of  $f$  to an index set  $I$ , i.e.,  $f_I = \sum_{m \in I} f_m$ . For a given set of indices  $I \subseteq \{1, \dots, M\}$ , let  $\kappa(I)$  be defined as follows:

$$\kappa(I) := \sup \left\{ \kappa \geq 0 \mid \kappa \leq \frac{\|\sum_{m \in I} f_m\|_{L_2(\Pi)}^2}{\sum_{m \in I} \|f_m\|_{L_2(\Pi)}^2}, \forall f_m \in \mathcal{H}_m (m \in I) \right\}.$$

$\kappa(I)$  represents the correlation of RKHSs inside the indices  $I$ . Similarly, we define the *canonical correlations* of RKHSs between  $I$  and  $I^c$  as follows:

$$\rho(I) := \sup \left\{ \frac{\langle f_I, g_{I^c} \rangle_{L_2(\Pi)}}{\|f_I\|_{L_2(\Pi)} \|g_{I^c}\|_{L_2(\Pi)}} \mid f_I \in \mathcal{H}_I, g_{I^c} \in \mathcal{H}_{I^c}, f_I \neq 0, g_{I^c} \neq 0 \right\}.$$

These quantities give a connection between the  $L_2(\Pi)$ -norm of  $f \in \mathcal{H}$  and the  $L_2(\Pi)$ -norm of  $\{f_m\}_{m \in I}$  as shown in the following lemma. The proof is given in Appendix B.

**Lemma 1** *For all  $I \subseteq \{1, \dots, M\}$ , we have*

$$\|f\|_{L_2(\Pi)}^2 \geq (1 - \rho(I)^2) \kappa(I) \left( \sum_{m \in I} \|f_m\|_{L_2(\Pi)}^2 \right).$$

We impose the following assumption for  $\kappa(I_0)$  and  $\rho(I_0)$ .

**Assumption 4 (Incoherence Assumption)** *For the truly active components  $I_0$ ,  $\kappa(I_0)$  is strictly positive and  $\rho(I_0)$  is strictly less than 1:*

$$(A4) \quad 0 < \kappa(I_0)(1 - \rho^2(I_0)).$$

This condition is known as the *incoherence condition* (Koltchinskii and Yuan, 2008, Meier et al., 2009), i.e., RKHSs are not too dependent on each other. In the theoretical analysis, we also obtain an upper bound of the  $L_2(\Pi)$ -norm of  $\hat{f} - f^*$  in terms of the  $L_2(\Pi)$ -norm of  $\{\hat{f}_m - f_m^*\}_{m \in I_0}$ . Thus, by the incoherence condition and Lemma 1, we may focus on bounding the  $L_2(\Pi)$ -norm of the “low-dimensional” components  $\{\hat{f}_m - f_m^*\}_{m \in I_0}$ , instead of all the components. Koltchinskii and Yuan (2010) considered a weaker condition including the *restricted isometry* (Candes and Tao, 2007) instead of (A4). Such a weaker condition is also applicable to our analysis, but we employ (A4) for simplicity.

Finally we impose the following technical assumption related to the sup-norm of the members in the RKHSs.

**Assumption 5 (Sup-norm Assumption)** *Along with the Spectral Assumption (A3), there exists a constant  $C_1$  such that*

$$(A5) \quad \|f_m\|_\infty \leq C_1 \|f_m\|_{L_2(\Pi)}^{1-s} \|f_m\|_{\mathcal{H}_m}^s \quad (\forall f_m \in \mathcal{H}_m, m = 1, \dots, M),$$

where  $s$  is the exponent defined in the Spectral Assumption (A3).

<sup>2</sup> The  $\epsilon$ -covering number  $\mathcal{N}(\epsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi))$  with respect to  $L_2(\Pi)$  is the minimal number of balls with radius  $\epsilon$  needed to cover the unit ball  $\mathcal{B}_{\mathcal{H}_m}$  in  $\mathcal{H}_m$  (van der Vaart and Wellner, 1996).

This assumption is satisfied if the RKHS is a Sobolev space or is continuously embeddable in a Sobolev space. For example, the RKHSs of Gaussian kernels are continuously embedded in all Sobolev spaces, and thus satisfy the Sup-norm Assumption (A5). More generally, RKHSs with  $m$ -times continuously differentiable kernels on a closed Euclidean ball in  $\mathbb{R}^d$  are also continuously embedded in a Sobolev space, and satisfy the Sup-norm Assumption (A5) with  $s = \frac{d}{2m}$  (see Corollary 4.36 of Steinwart (2008)). Therefore this assumption is somewhat common for practically used kernels. A more general necessary and sufficient condition in terms of *real interpolation* is shown in Bennett and Sharpley (1988). Steinwart et al. (2009) used this assumption to show the optimal rates for regularized regression using a single kernel function, and one can find detailed discussions about the assumption there.

### 3 Convergence rate analysis

In this section, we present our main result.

#### 3.1 The convergence rate of elastic-net MKL

Here we derive the learning rate of the estimator  $\hat{f}$  defined by Eq. (1). We denote the number of truly active components by  $d := |I_0|$ . We may suppose that the number of kernels  $M$  and the number of active kernels  $d$  are increasing with respect to the number of samples  $n$ . Our main purpose of this section is to show that the learning rate can be faster than the existing bounds. The existing bound has already been shown to be optimal on the  $\ell_\infty$ -mixed-norm ball Koltchinskii and Yuan (2010), Raskutti et al. (2010). Our claim is that the convergence rate can further achieve the minimax optimal rate on the  $\ell_2$ -mixed-norm ball, which is faster than that on the  $\ell_\infty$ -mixed-norm ball.

Define  $\eta(t)$  for  $t > 0$  as

$$\eta(t) := \max(1, \sqrt{t}, t/\sqrt{n}).$$

For given  $\lambda > 0$ , we define  $\xi_n$  as

$$\xi_n := \xi_n(\lambda) = \left( \frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}} \vee \sqrt{\frac{\log(M)}{n}} \right). \quad (4)$$

**Theorem 2** Suppose Assumptions 1–5 are satisfied, and let  $\lambda > 0$  be an arbitrary positive number. Then there exist universal constants  $\tilde{C}_1, \tilde{C}_2$  and a constant  $\psi_s$  depending on  $s, c, L, C_1$  such that if  $\lambda_1^{(n)}, \lambda_2^{(n)}$  and  $\lambda_3^{(n)}$  are set as  $\lambda_1^{(n)} = \psi_s \eta(t) \xi_n(\lambda)$ ,  $\lambda_2^{(n)} = \lambda$ ,  $\lambda_3^{(n)} = \lambda$ , then for all  $n$  and  $r(> 0)$  satisfying  $\frac{\log(M)}{\sqrt{n}} \leq 1$  and the inequality

$$\frac{\tilde{C}_1 \max(\psi_s \sqrt{n} \xi_n^2, r) \left( d + \frac{\lambda_3^{(n)1+q}}{\lambda_1^{(n)2}} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right)}{(1 - \rho(I_0)^2) \kappa(I_0)} \leq 1,$$

we have

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{\tilde{C}_2}{(1 - \rho(I_0))^2 \kappa(I_0)} \left( d \lambda_1^{(n)2} + \lambda_3^{(n)1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right), \quad (5)$$

with probability  $1 - \exp(-t) - \exp\left(-\min\left\{\frac{r^2 \log(M)}{n \xi_n(\lambda)^4 \psi_s^2}, \frac{r}{\xi_n(\lambda)^2 \psi_s}\right\}\right)$  for all  $t \geq 1$ .

A proof of Theorem 2 is provided in Appendix D. The convergence rate (5) contains a tuning parameter  $\lambda$ . Here we optimize this parameter. Let

$$R_{p,g^*} := \left( \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^p \right)^{\frac{1}{p}},$$

and we assume that  $R_{p,g^*}$  is strictly positive for all  $p \geq 1$  ( $R_{p,g^*} > 0$ ). If  $n$  is sufficiently large compared with  $R_{2,g^*}$ , the RHS of Eq. (5) is minimized by

$$\lambda = d^{\frac{1}{1+q+s}} n^{-\frac{1}{1+q+s}} R_{2,g^*}^{-\frac{2}{1+q+s}},$$

up to constants. Then the convergence rate (5) is reduced to

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \hat{C}_1 \left( d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d \log(M)}{n} + d^{\frac{q+s}{1+q+s}} n^{-\frac{1+q}{1+q+s} - \frac{q(1-s)}{(1+s)(1+q+s)}} R_{2,g^*}^{\frac{2}{1+q+s}} \right), \quad (6)$$

where  $\widehat{C}_1$  is a constant. If  $n^{\frac{q}{1+q}} \frac{d}{R_{2,g^*}^2} \geq C$  with a constant  $C$  (this holds if  $\|g_m^*\|_{\mathcal{H}_m} \leq \sqrt{C}$  for all  $m$ ), then Eq. (6) becomes

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \widehat{C}_2 \left( d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d \log(M)}{n} \right), \quad (7)$$

where  $\widehat{C}_2$  is a constant. We see that, as  $q$  becomes large (the truth becomes smooth) or  $s$  becomes small (the RKHSs become simple), the convergence rate becomes faster when  $R_{2,g^*} \geq 1$ . In the next subsection, we show that this bound (7) achieves the minimax optimal rate on the  $\ell_2$ -mixed-norm ball.

### 3.2 Minimax learning rate of $\ell_2$ -mixed-norm ball

To derive the minimax rate, we slightly simplify the setup. First, we assume that the input  $\mathcal{X}$  is expressed as  $\mathcal{X} = \tilde{\mathcal{X}}^M$  for some space  $\tilde{\mathcal{X}}$ . Second, all the RKHSs  $\{\mathcal{H}_m\}_{m=1}^M$  are the same as an RKHS  $\tilde{\mathcal{H}}$  defined on  $\tilde{\mathcal{X}}$ . Finally, we assume that the marginal distribution  $\Pi$  of input is a product of a probability distribution  $Q$ , i.e.,  $\Pi = Q^M$ . Thus, an input  $x = (\tilde{x}^{(1)}, \dots, \tilde{x}^{(M)}) \in \mathcal{X} = \tilde{\mathcal{X}}^M$  is a concatenation of  $M$  random variables  $\{\tilde{x}^{(m)}\}_{m=1}^M$  independently and identically distributed from the distribution  $Q$ . Moreover, the function class  $\mathcal{H}$  is a class of functions  $f$  such that

$$f(x) = f(\tilde{x}^{(1)}, \dots, \tilde{x}^{(M)}) = \sum_{m=1}^M f_m(\tilde{x}^{(m)}),$$

where  $f_m \in \tilde{\mathcal{H}}$  for all  $m$ . Without loss of generality, we may assume that all functions in  $\tilde{\mathcal{H}}$  are centered:

$$\mathbb{E}_{\tilde{X} \sim Q}[f(\tilde{X})] = 0 \quad (\forall f \in \tilde{\mathcal{H}}).$$

We assume that the spectrum of the kernel  $\tilde{k}$  corresponding to the RKHS  $\tilde{\mathcal{H}}$  decays at the rate of  $-\frac{1}{s}$ . That is, in addition to Assumption 3, we impose the following lower bound to the spectrum: there exist  $c', c (> 0)$  such that

$$c' k^{-\frac{1}{s}} \leq \mu_k \leq c k^{-\frac{1}{s}}, \quad (8)$$

where  $\{\mu_k\}_k$  is the spectrum of the kernel  $\tilde{k}$  (see Eq.(2)). We also assume that the noise  $\{\epsilon_i\}_{i=1}^n$  is generated by a Gaussian distribution with mean 0 and standard deviation  $\sigma$ .

Let  $\mathcal{H}_{\ell_0}(d)$  be the set of functions with  $d$  non-zero components in  $\mathcal{H}$  defined by

$$\mathcal{H}_{\ell_0}(d) := \{(f_1, \dots, f_M) \in \mathcal{H} \mid |\{m \mid \|f_m\|_{\mathcal{H}_m} \neq 0\}| \leq d\}.$$

We define  $\ell_p$ -mixed-norm ball ( $p \geq 1$ ) with radius  $R$  in  $\mathcal{H}_0(d)$  as

$$\mathcal{H}_{\ell_p}^{d,q}(R) := \left\{ f = \sum_{m=1}^M f_m \mid \exists (g_1, \dots, g_M) \in \mathcal{H}_0(d), f_m = T_m^{\frac{q}{2}} g_m, \left( \sum_{m=1}^M \|g_m\|_{\mathcal{H}_m}^p \right)^{\frac{1}{p}} \leq R \right\}.$$

In Raskutti et al. (2010), the minimax learning rate on  $\mathcal{H}_{\ell_\infty}^{d,0}(R)$  (i.e.,  $p = \infty$  and  $q = 0$ ) was derived<sup>3</sup>. We show (a lower bound of) the minimax learning rate for more general settings ( $p = 2, \infty$  and  $0 \leq q \leq 1$ ) in the following theorem.

**Theorem 3** *Let  $\tilde{s} = \frac{s}{1+q}$ . Assume  $d \leq M/4$ . Then the minimax learning rates are lower bounded as follows. There exists a constant  $\tilde{C}_1$  such that for  $R_2 \geq \sqrt{\frac{d \log(M/d)}{n}}$ , the radius of the  $\ell_2$ -mixed-norm ball, we have*

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_2}^{d,q}(R_2)} \mathbb{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] \geq \tilde{C}_1 \left( d^{\frac{1}{1+\tilde{s}}} n^{-\frac{1}{1+\tilde{s}}} R_2^{\frac{2\tilde{s}}{1+\tilde{s}}} + \frac{d \log(M/d)}{n} \right), \quad (9)$$

where ‘inf’ is taken over all measurable functions of the samples  $(x_i, y_i)_{i=1}^n$  and the expectation is taken for the sample distribution. Similarly, we have the following minimax-rate for  $p = \infty$ :

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R_\infty)} \mathbb{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] \geq \tilde{C}_1 \left( d n^{-\frac{1}{1+\tilde{s}}} R_\infty^{\frac{2\tilde{s}}{1+\tilde{s}}} + \frac{d \log(M/d)}{n} \right), \quad (10)$$

for  $R_\infty \geq \sqrt{\frac{\log(M/d)}{n}}$ .

<sup>3</sup> The set  $\mathcal{F}_{M,d,\mathcal{H}}(R)$  in Raskutti et al. (2010) corresponds to  $\mathcal{H}_{\ell_\infty}^{d,0}(R)$  in the current paper.

A proof of Theorem 3 is provided in Appendix E.

Obviously, our learning rate (7) of elastic-net MKL achieves the minimax optimal rate (9) on the  $\ell_2$ -mixed-norm ball if  $M \gg d$ . Moreover, the optimal rate (9) on the  $\ell_2$ -mixed-norm ball is always faster than that of  $\ell_\infty$ -mixed-norm (10). To see this, let  $R_{\infty, g^*} := \max_m \|g_m^*\|_{\mathcal{H}_m}$ ; then we always have  $R_{2, g^*} \leq \sqrt{d} R_{\infty, g^*}$  and consequently we have

$$d^{\frac{1}{1+s}} n^{-\frac{1}{1+s}} R_{2, g^*}^{\frac{2s}{1+s}} \leq d n^{-\frac{1}{1+s}} R_{\infty, g^*}^{\frac{2s}{1+s}}.$$

Now we consider two examples, “inhomogeneous setting” and “homogeneous setting”, to compare these two bounds:

1.  $\|g_m^*\|_{\mathcal{H}_m} = m^{-1}$  ( $\forall m \in I_0 = \{1, \dots, d\}$ ) (inhomogeneous setting): In this situation,  $R_{\infty, g^*} = 1$  and  $R_{2, g^*} \leq 1$ . Thus, the learning rate (7) of elastic-net MKL and the minimax rate on the  $\ell_2$ -mixed-norm ball are  $d^{\frac{1}{1+s}} n^{-\frac{1}{1+s}} + \frac{d \log(M)}{n}$  and that on the  $\ell_\infty$ -mixed-norm ball is  $d n^{-\frac{1}{1+s}} + \frac{d \log(M)}{n}$ . Therefore, in the first term (the leading term with respect to  $n$ ), there is a difference in the  $d^{\frac{s}{1+s}}$  factor. This difference could be  $\sqrt{d}$  in the worst case. Thus, there appears large discrepancy between the two rates in high-dimensional settings.
2.  $\|g_m^*\|_{\mathcal{H}_m} = 1$  ( $\forall m \in I_0$ ) (homogeneous setting): In this situation,  $R_{\infty, g^*} = 1$  and  $R_{2, g^*} = \sqrt{d}$ . Thus, all the bounds are  $d n^{-\frac{1}{1+s}} + \frac{d \log(M)}{n}$ . Here we observe that the learning rate (7) of elastic-net MKL coincides with the minimax rate on the  $\ell_\infty$ -mixed-norm ball. We also notice that the homogeneous setting is the only situation where those two rates coincide with each other. As seen later, the existing bounds by previous works are the minimax rate on the  $\ell_\infty$ -mixed-norm ball, thus are tight only in the homogeneous setting.

### 3.3 Comparison with existing bounds

Here we compare the existing bounds and the bound we derived. Roughly speaking, the difference from the existing bounds is summarized in the following two points:

- (a) Our learning rate achieves the minimax-rate of  $\ell_2$ -mixed-norm ball, instead of the  $\ell_\infty$ -mixed-norm ball.
- (b) Our bound includes the smoothing parameter  $q$  (Assumption 2), and thus is more general and faster than existing bounds.

The first bound on the convergence rate of MKL was derived by Koltchinskii and Yuan (2008), which assumed  $q = 1$  and  $\frac{1}{d} \sum_{m \in I_0} \frac{\|g_m^*\|_{\mathcal{H}_m}^2}{\|f_m^*\|_{\mathcal{H}_m}^2} \leq C$ . Under these rather strong conditions, they showed the bound  $d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} + \frac{d \log(M)}{n}$ . For the smooth case  $q = 1$ , we obtained a faster rate  $n^{-\frac{2}{2+s}}$  instead of  $n^{-\frac{1}{1+s}}$  in their bound with respect to  $n$ .

The second bound was given by Meier et al. (2009), which showed  $d \left( \frac{\log(M)}{n} \right)^{\frac{1}{1+s}}$  for elastic-net regularization (1) under  $q = 0$ . Their bound almost achieves the minimax rate on the  $\ell_\infty$ -mixed-norm ball except the additional  $\log(M)$  term. Compared with our bound, their bound has the  $\log(M)$  term and the rate with respect to  $d$  is larger than  $d^{\frac{1}{1+s}}$  in our bound.

Most recently, Koltchinskii and Yuan (2010) presented the bound  $n^{-\frac{1}{1+s}} (d + \sum_{m \in I_0} \|f_m^*\|_{\mathcal{H}_m}) + \frac{d \log(M)}{n}$  for  $q = 0$ . Their bound is exactly the minimax rate on the  $\ell_\infty$ -mixed-norm ball. However, their bound is  $d^{\frac{s}{1+s}}$  times slower than ours if the ground truth is inhomogeneous. For example, when  $\|f_m^*\|_{\mathcal{H}_m} = m^{-1}$  ( $m \in I_0 = \{1, \dots, d\}$ ) and  $f_m^* = 0$  (otherwise), their bound is  $n^{-\frac{1}{1+s}} d + \frac{d \log(M)}{n}$ , while our bound is  $n^{-\frac{1}{1+s}} d^{\frac{1}{1+s}} + \frac{d \log(M)}{n}$ .

All the bounds explained above focused on either  $q = 0$  or 1. On the other hand, our analysis is more general in that the whole range of  $0 \leq q \leq 1$  can be accommodated.

The relation between our analysis and existing analyses are summarized in Table 3.3.

## 4 Conclusion and Discussion

We presented a new learning rate of elastic-net MKL, which is faster than the existing bounds of several MKL formulations. According to our bound, the learning rate of elastic-net MKL achieves the minimax rate on the  $\ell_2$ -mixed-norm ball, instead of the  $\ell_\infty$ -mixed-norm ball. Our bound includes

Table 1: Relation between our analysis and existing analyses.

	regularizer	smoothness ( $q$ )	minimaxity	convergence rate
K&Y (2008)	$\ell_1$	$q = 1$	?	$d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} + \frac{d \log(M)}{n^{\frac{1}{1+s}}}$
Meier et al. (2009)	Elastic-net	$q = 0$	not optimal	$d \left( \frac{\log(M)}{n} \right)^{\frac{1}{1+s}}$
K&Y (2010)	variant of $\ell_1$	$q = 0$	$\ell_\infty$ -ball	$dn^{-\frac{1}{1+s}} + \frac{d \log(M)}{n^{\frac{1}{1+s}}}$
This paper	Elastic-net	$0 \leq q \leq 1$	$\ell_2$ -ball	$\left(\frac{d}{n}\right)^{\frac{1+q}{1+q+s}} R_{2,q^*}^{\frac{2s}{1+q+s}} + \frac{d \log(M)}{n}$

a parameter  $s$  representing the complexity of the RKHSs and another parameter  $q$  controlling the smoothness of the truth. Under a natural condition, the learning rate becomes faster as  $s$  becomes small or  $q$  becomes large. Although the existing works concluded that MKL is optimal in a sense that it achieves the minimax rate of the  $\ell_\infty$ -mixed-norm ball, we presented that elastic-net MKL further achieves the minimax rate of the  $\ell_2$ -mixed-norm ball which is faster than that of the  $\ell_\infty$ -mixed-norm ball.

Koltchinskii and Yuan (2010) considered a variant of  $\ell_1$  regularization:  $\sum_{m=1}^M \lambda_1^{(n)} \|f_m\|_n + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}$ . They showed that MKL with that regularization achieves the minimax rate of the  $\ell_\infty$ -mixed-norm ball. It might be interesting to investigate whether that regularization also achieves the minimax rate of the  $\ell_2$ -mixed-norm ball or another faster rate. In particular, it is interesting to study whether the smoothness parameterization ( $q$ ) gives a faster rate also for that  $\ell_1$  regularization. If not, that might explain the effectiveness of the elastic-net regularization in real data experiments.

## A Covering Number

Here, we give a detailed characterization of the covering number in terms of the spectrum using the operator  $T_m$ . Accordingly, we give the complexity of the set of functions satisfying the Convolution Assumption (Assumption 2). We extend the domain and the range of the operator  $T_m$  to the whole space of  $L_2(\Pi)$ , and define its power  $T_m^\beta : L_2(\Pi) \rightarrow L_2(\Pi)$  for  $\beta \in [0, 1]$  as

$$T_m^\beta f := \sum_{k=1}^{\infty} \mu_{k,m}^\beta \langle f, \phi_{k,m} \rangle_{L_2(\Pi)} \phi_{k,m}, \quad (f \in L_2(\Pi)).$$

Moreover, we define a Hilbert space  $\mathcal{H}_{m,\beta}$  as

$$\mathcal{H}_{m,\beta} := \left\{ \sum_{k=1}^{\infty} b_k \phi_{k,m} \mid \sum_{k=1}^{\infty} \mu_{k,m}^{-\beta} b_k^2 \leq \infty \right\},$$

and equip this space with the Hilbert space norm  $\|\sum_{k=1}^{\infty} b_k \phi_{k,m}\|_{\mathcal{H}_{m,\beta}} := \sqrt{\sum_{k=1}^{\infty} \mu_{k,m}^{-\beta} b_k^2}$ . One can check that  $\mathcal{H}_{m,1} = \mathcal{H}_m$ . Here we define, for  $R > 0$ ,

$$\mathcal{H}_m^q(R) := \{f_m = T_m^{\frac{q}{2}} g_m \mid g_m \in \mathcal{H}_m, \|g_m\|_{\mathcal{H}_m} \leq R\}. \quad (11)$$

Then we obtain the following lemma.

**Lemma 4**  $\mathcal{H}_m^q(1)$  is equivalent to the unit ball of  $\mathcal{H}_{m,1+q}$ :  $\mathcal{H}_m^q(1) = \{f_m \in \mathcal{H}_{m,1+q} \mid \|f_m\|_{\mathcal{H}_m} \leq 1\}$ .

This can be shown as follows. For all  $f_m \in \mathcal{H}_m^q(1)$ , there exists  $g_m \in \mathcal{H}_m$  such that  $f_m = T_m^{\frac{q}{2}} g_m$  and  $\|g_m\|_{\mathcal{H}_m} \leq 1$ . Thus,  $g_m = (T_m^{\frac{q}{2}})^{-1} f_m = \sum_{k=1}^{\infty} \mu_{k,m}^{-\frac{q}{2}} \langle f, \phi_{k,m} \rangle_{L_2(\Pi)} \phi_{k,m}$  and  $1 \geq \|g_m\|_{\mathcal{H}_m} = \sum_{k=1}^{\infty} \mu_{k,m}^{-1} \langle g, \phi_{k,m} \rangle_{L_2(\Pi)}^2 = \sum_{k=1}^{\infty} \mu_{k,m}^{-(1+q)} \langle f, \phi_{k,m} \rangle_{L_2(\Pi)}^2$ . Therefore,  $f \in \mathcal{H}_m$  is in  $\mathcal{H}_m^q(1)$  if and only if the norm of  $f$  in  $\mathcal{H}_{m,1+q}$  is well-defined and not greater than 1.

Now Theorem 15 of Steinwart et al. (2009) gives an upper bound of the covering number of the unit ball  $\mathcal{B}_{\mathcal{H}_{m,\beta}}$  in  $\mathcal{H}_{m,\beta}$  as  $\mathcal{N}(\varepsilon, \mathcal{B}_{\mathcal{H}_{m,\beta}}, L_2(\Pi)) \leq C \varepsilon^{-2\frac{s}{1+s}}$ , where  $C$  is a constant depending on  $c, s, \beta$ . This inequality with  $\beta = 1$  corresponds to Eq. (3). Moreover, substituting  $\beta = 1 + q$  into the above equation, we have

$$\mathcal{N}(\varepsilon, \mathcal{H}_m^q(1), L_2(\Pi)) \leq C \varepsilon^{-2\frac{s}{1+q}}. \quad (12)$$



## B Proof of Lemma 1

**Proof: (Lemma 1)** For  $J = I^c$ , we have

$$\begin{aligned} Pf^2 &= \|f_I\|_{L_2(\Pi)}^2 + 2\langle f_I, f_J \rangle_{L_2(\Pi)} + \|f_J\|_{L_2(\Pi)}^2 \geq \|f_I\|_{L_2(\Pi)}^2 - 2\rho(I)\|f_I\|_{L_2(\Pi)}\|f_J\|_{L_2(\Pi)} + \|f_J\|_{L_2(\Pi)}^2 \\ &\geq (1 - \rho(I)^2)\|f_I\|_{L_2(\Pi)}^2 \geq (1 - \rho(I)^2)\kappa(I) \left( \sum_{m \in I} \|f_m\|_{L_2(\Pi)}^2 \right), \end{aligned}$$

where we used the inequality of arithmetic and geometric mean in the second inequality.  $\blacksquare$

## C Talagrand's Concentration Inequality

**Proposition 5 (Talagrand's Concentration Inequality (Talagrand, 1996, Bousquet, 2002))** *Let  $\mathcal{G}$  be a function class on  $\mathcal{X}$  that is separable with respect to  $\infty$ -norm, and  $\{x_i\}_{i=1}^n$  be i.i.d. random variables with values in  $\mathcal{X}$ . Furthermore, let  $B \geq 0$  and  $U \geq 0$  be  $B := \sup_{g \in \mathcal{G}} \mathbb{E}[(g - \mathbb{E}[g])^2]$  and  $U := \sup_{g \in \mathcal{G}} \|g\|_\infty$ , then there exists a universal constant  $K$  such that, for  $Z := \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(x_i) - \mathbb{E}[g] \right|$ , we have*

$$P \left( Z \geq K \left[ \mathbb{E}[Z] + \sqrt{\frac{Bt}{n}} + \frac{Ut}{n} \right] \right) \leq e^{-t},$$

for all  $t > 0$ .

## D Proof of Theorem 2

For a Hilbert space  $\mathcal{G} \subset L_2(P)$ , let the  $i$ -th entropy number  $e_i(\mathcal{G} \rightarrow L(P))$  be the infimum of  $\epsilon > 0$  for which  $\mathcal{N}(\epsilon, \mathcal{B}_{\mathcal{G}}, L_2(P)) \leq 2^{i-1}$ , where  $\mathcal{B}_{\mathcal{G}}$  is the unit ball of  $\mathcal{G}$ . One can check that if the spectral assumption (A3) holds, the  $i$ -th entropy number is bounded as

$$e_i(\mathcal{H}_m \rightarrow L_2(\Pi)) \leq \tilde{c} i^{-\frac{1}{2s}}. \quad (13)$$

where  $\tilde{c}$  is a constant depends on  $s$  and  $c$ .

The following proposition is the key of the localization.

**Proposition 6** *Let  $\mathcal{B}_{\sigma,a,b} \subset \mathcal{H}_m$  be a set such that  $\mathcal{B}_{\sigma,a,b} = \{f_m \in \mathcal{H}_m \mid \|f_m\|_{L_2(\Pi)} \leq \sigma, \|f_m\|_{\mathcal{H}_m} \leq a, \|f_m\|_\infty \leq b\}$ . Assume the Spectral Assumption (A3), then there exist constants  $\tilde{c}_s, C'_s$  depending only on  $s$  and  $c$  such that*

$$\mathbb{E} \left[ \sup_{f_m \in \mathcal{B}_{\sigma,a,b}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i) \right| \right] \leq C'_s \left( \frac{\sigma^{1-s} (\tilde{c}_s a)^s}{\sqrt{n}} \vee (\tilde{c}_s a)^{\frac{2s}{1+s}} b^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} \right).$$

**Proof: (Proposition 6)** Let  $D_n$  be the empirical distribution:  $D_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ . To bound empirical processes, a bound of the entropy number with respect to the empirical  $L_2$ -norm is needed. Corollary 7.31 of Steinwart (2008) gives the following upper bound: under the condition (13), there exists a constant  $c_s > 0$  only depending on  $s$  such that

$$\mathbb{E}_{D_n \sim \Pi^n} [e_i(\mathcal{H}_m \rightarrow L_2(D_n))] \leq c_s \tilde{c} i^{-\frac{1}{2s}}.$$

Finally this and Theorem 7.16 of Steinwart (2008) gives the assertion.  $\blacksquare$

Using the above proposition and the *peeling device*, we obtain the following lemma (see also Meier et al. (2009)).

**Lemma 7** *Under the Spectral Assumption (Assumption 3), there exists a constant  $C_s$  depending only on  $s$  and  $C$  such that for all  $\lambda > 0$*

$$\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m : \|f_m\|_{\mathcal{H}_m} \leq 1} \frac{\left| \frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i) \right|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda}} \right] \leq C_s \left( \frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}} \right).$$

**Proof: (Lemma 7)** Let  $\mathcal{H}_m(\sigma) := \{f_m \in \mathcal{H}_m \mid \|f_m\|_{\mathcal{H}_m} \leq 1, \|f_m\|_{L_2(\Pi)} \leq \sigma\}$  and  $z = 2^{1/s} > 1$ . Then by noticing  $\|f_m\|_\infty \leq \|f_m\|_{\mathcal{H}_m}$ , Proposition 6 gives

$$\begin{aligned}
& \mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m: \|f_m\|_{\mathcal{H}_m} \leq 1} \frac{|\frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda}} \right] \\
& \leq \mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m(\lambda^{1/2})} \frac{|\frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda}} \right] + \sum_{k=1}^{\infty} \mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m(z^k \lambda^{1/2}) \setminus \mathcal{H}_m(z^{k-1} \lambda^{1/2})} \frac{|\frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda}} \right] \\
& \leq C'_s \left( \frac{\lambda^{\frac{1-s}{2}} \tilde{c}_s^s}{\lambda^{\frac{1}{2}} \sqrt{n}} \vee \frac{\tilde{c}_s^{\frac{2s}{1+s}}}{n^{\frac{1}{1+s}} \lambda^{\frac{1}{2}}} \right) + \sum_{k=0}^{\infty} C'_s \left( \frac{z^{k(1-s)} \lambda^{\frac{1-s}{2}} \tilde{c}_s^s}{\sqrt{n} z^k \lambda^{\frac{1}{2}}} \vee \frac{\tilde{c}_s^{\frac{2s}{1+s}}}{n^{\frac{1}{1+s}} z^k \lambda^{\frac{1}{2}}} \right) \\
& = C'_s \left( \tilde{c}_s^s \sqrt{\frac{\lambda^{-s}}{n}} \vee \tilde{c}_s^{\frac{2s}{1+s}} \left( \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}} \right) \right) + \sum_{k=0}^{\infty} C'_s \left( \tilde{c}_s^s z^{-sk} \sqrt{\frac{\lambda^{-s}}{n}} \vee \tilde{c}_s^{\frac{2s}{1+s}} z^{-k} \left( \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}} \right) \right) \\
& \leq 2C'_s \left( \frac{1}{1-z^{-s}} \tilde{c}_s^s \sqrt{\frac{\lambda^{-s}}{n}} + \frac{1}{1-z^{-1}} \tilde{c}_s^{\frac{2s}{1+s}} \left( \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}} \right) \right) = 2C'_s \left( 2\tilde{c}_s^s \sqrt{\frac{\lambda^{-s}}{n}} + \frac{2^{1/s}}{2^{1/s}-1} \tilde{c}_s^{\frac{2s}{1+s}} \left( \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}} \right) \right) \\
& \leq 2C'_s \left( 2\tilde{c}_s^s + \frac{2^{1/s}}{2^{1/s}-1} \tilde{c}_s^{\frac{2s}{1+s}} \right) \left( \sqrt{\frac{\lambda^{-s}}{n}} \vee \left( \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}} \right) \right).
\end{aligned}$$

By setting  $C_s \leftarrow 2C'_s \left( 2\tilde{c}_s^s + \frac{2^{1/s}}{2^{1/s}-1} \tilde{c}_s^{\frac{2s}{1+s}} \right)$ , we obtain the assertion.  $\blacksquare$

The above lemma immediately gives the following corollary.

**Corollary 8** *Under the Spectral Assumption (Assumption 3), for all  $\lambda > 0$*

$$\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right] \leq C_s \left( \frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}} \right),$$

where  $C_s$  is the constant appeared in the statement of Lemma 7, and we employed a convention such that  $\frac{0}{0} = 0$ .

Moreover we obtain the following corollary.

**Corollary 9** *Under the Spectral Assumption (Assumption 3), for all  $\lambda > 0$*

$$\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right] \leq 2C_s L \left( \frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}} \right),$$

where  $C_s$  is the constant appeared in the statement of Lemma 7.

**Proof: (Corollary 9)** Here we write  $Pf = \mathbb{E}[f]$  and  $P_n f = \frac{1}{n} \sum_{i=1}^n f(x_i, y_i)$  for a function  $f$ . Notice that  $P\epsilon f_m = 0$ , thus  $\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i) = (P_n - P)(\epsilon f_m)$ . By the symmetrization argument (van der Vaart and Wellner, 1996, Lemma 2.3.1) and the contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12), we obtain

$$\begin{aligned}
\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \frac{|(P - P_n)(\epsilon f_m)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right] &= \mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \left| (P - P_n) \frac{\epsilon f_m}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right| \right] \\
&\leq 2\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \left| \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i \epsilon_i f_m(x_i)}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right| \right] \\
&\leq 2L\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \left| \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i)}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right| \right]
\end{aligned}$$

$$\leq 2C_s L \left( \frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}} \right).$$

This gives the assertion. ■

From now on, we refer to  $C_s$  as the constant appeared in the statement of Lemma 7. We define  $\tilde{\phi}_s$  as

$$\tilde{\phi}_s = 2KL(C_s + 1 + C_1).$$

Remind the definition of  $\xi_n$  (Eq. (4)), then we obtain the following theorem.

**Theorem 10** *Under the Basic Assumption, the Spectral Assumption and the Supnorm Assumption, when  $\frac{\log(M)}{\sqrt{n}} \leq 1$ , we have for all  $\lambda > 0$  and all  $t \geq 1$*

$$\left| \frac{1}{n} \sum_{i=1}^n \epsilon_i (\hat{f}_m(x_i) - f_m^*(x_i)) \right| \leq \tilde{\phi}_s \xi_n(\lambda) \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \max(1, \sqrt{t}, t/\sqrt{n}),$$

$$(\forall f_m \in \mathcal{H}_m, \forall m = 1, \dots, M),$$

with probability  $1 - \exp(-t)$ . Moreover we also have

$$\mathbb{E} \left[ \max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right] \leq 4\tilde{\phi}_s \xi_n.$$

**Proof: (Theorem 10)** Since

$$\frac{\|f_m\|_{L_2(\Pi)}}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \leq 1,$$

$$\frac{\|f_m\|_{\infty}}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \leq \frac{C_1 \|f_m\|_{L_2(\Pi)}^{1-s} \|f_m\|_{\mathcal{H}_m}^s}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \stackrel{\text{Young}}{\leq} \frac{C_1 \lambda^{-\frac{s}{2}} \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}}$$

$$\leq C_1 \lambda^{-\frac{s}{2}},$$

applying Talagrand's concentration inequality (Proposition 5), we obtain

$$P \left( \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \geq K \left[ 2C_s L \xi_n + \sqrt{\frac{L^2 t}{n}} + \frac{C_1 L \lambda^{-\frac{s}{2}} t}{n} \right] \right) \leq e^{-t}.$$

Therefore the uniform bound over all  $m = 1, \dots, M$  is given as

$$P \left( \max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \geq K \left[ 2C_s L \xi_n + \sqrt{\frac{L^2 t}{n}} + \frac{C_1 L \lambda^{-\frac{s}{2}} t}{n} \right] \right)$$

$$\leq \sum_{m=1}^M P \left( \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \geq K \left[ 2C_s L \xi_n + \sqrt{\frac{L^2 t}{n}} + \frac{C_1 L \lambda^{-\frac{s}{2}} t}{n} \right] \right)$$

$$\leq M e^{-t}.$$

Setting  $t \leftarrow t + \log(M)$ , we have

$$P \left( \max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \geq K \left[ 2C_s L \xi_n + \sqrt{\frac{L^2(t + \log(M))}{n}} + \frac{C_1 L \lambda^{-\frac{s}{2}}(t + \log(M))}{n} \right] \right) \leq e^{-t}. \quad (14)$$

Now

$$\sqrt{\frac{L^2(t + \log(M))}{n}} + \frac{C_1 L \lambda^{-\frac{s}{2}}(t + \log(M))}{n} \leq L \sqrt{\frac{t}{n}} + L \sqrt{\frac{\log(M)}{n}} + \frac{C_1 L \lambda^{-\frac{s}{2}}}{\sqrt{n}} \left( \frac{t}{\sqrt{n}} + \frac{\log(M)}{\sqrt{n}} \right)$$

$$\leq \xi_n \left( L\sqrt{t} + L + C_1 L \frac{t}{\sqrt{n}} + C_1 L \right) \leq \xi_n (2L + 2C_1 L) \eta(t).$$

where we used  $\frac{\log(M)}{\sqrt{n}} \leq 1$  in the second inequality. Thus Eq. (14) implies

$$P \left( \max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \geq K(2C_s L + 2L + 2C_1 L) \xi_n \eta(t) \right) \leq e^{-t}.$$

By substituting  $\tilde{\phi}_s = 2KL(C_s + 1 + C_1)$ , we obtain

$$P \left( \max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \geq \tilde{\phi}_s \xi_n \eta(t) \right) \leq e^{-t}, \quad (15)$$

which gives the first assertion.

Next we show the second assertion. Eq. (15) implies that

$$\begin{aligned} \mathbb{E} \left[ \max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right] &\leq \tilde{\phi}_s \xi_n + \sum_{t=0}^{\infty} e^{-t} \tilde{\phi}_s \xi_n \eta(t+1) \\ &\leq \tilde{\phi}_s \xi_n + \tilde{\phi}_s \xi_n \sum_{t=0}^{\infty} e^{-t} (t+1) \leq 4\tilde{\phi}_s \xi_n, \end{aligned}$$

where we used  $\eta(t+1) = \max\{1, \sqrt{t+1}, (t+1)/\sqrt{n}\} \leq t+1$  in the second inequality. Thus we obtain the assertion.  $\blacksquare$

Moreover we obtain the following bound for the difference of the empirical and the expectation  $L_2$ -norm. Let  $\tilde{\phi}'_s$  be

$$\tilde{\phi}'_s = K [16KC_1(C_s + 1 + C_1) + C_1 + C_1^2].$$

We define  $\zeta_n(r, \lambda)$  as

$$\zeta_n(r, \lambda) := \min \left( \frac{r^2 \log(M)}{n \xi_n(\lambda)^4 \tilde{\phi}_s'^2}, \frac{r}{\xi_n(\lambda)^2 \tilde{\phi}_s'} \right).$$

**Theorem 11** *Under the Spectral Assumption and the Supnorm Assumption, when  $\frac{\log(M)}{\sqrt{n}} \leq 1$ , for all  $\lambda > 0$  we have*

$$\left| \left\| \sum_{m=1}^M f_m \right\|_n^2 - \left\| \sum_{m=1}^M f_m \right\|_{L_2(\Pi)}^2 \right| \leq \max(\tilde{\phi}'_s \sqrt{n} \xi_n^2(\lambda), r) \left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2,$$

( $\forall f_m \in \mathcal{H}_m$  ( $m = 1, \dots, M$ )),

with probability  $1 - \exp(-\zeta_n(r, \lambda))$ .

**Proof: (Theorem 11)**

$$\begin{aligned} &\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \frac{\left| \left\| \sum_{m=1}^M f_m \right\|_n^2 - \left\| \sum_{m=1}^M f_m \right\|_{L_2(\Pi)}^2 \right|}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \right] \\ &\leq 2\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \frac{\left| \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \sum_{m=1}^M f_m(x_i) \right)^2 \right|}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \right] \\ &\leq \sup_{f_m \in \mathcal{H}_m} \frac{\left\| \sum_{m=1}^M f_m \right\|_{\infty}}{\sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \times 2\mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \frac{\left| \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \sum_{m=1}^M f_m(x_i) \right) \right|}{\sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right], \quad (16) \end{aligned}$$

where we used the contraction inequality in the last line (Ledoux and Talagrand, 1991, Theorem 4.12). Here we notice that

$$\begin{aligned} \left\| \sum_{m=1}^M f_m \right\|_{\infty} &\leq \sum_{m=1}^M C_1 \|f_m\|_{L_2(\Pi)}^{1-s} \|f_m\|_{\mathcal{H}_m}^s \leq \sum_{m=1}^M C_1 \lambda^{-\frac{s}{2}} \sqrt{\|f_m\|_{L_2(\Pi)}^{2(1-s)} (\lambda \|f_m\|_{\mathcal{H}_m}^2)^s} \\ &\leq \sum_{m=1}^M C_1 \lambda^{-\frac{s}{2}} \sqrt{(1-s) \|f_m\|_{L_2(\Pi)}^2 + s \lambda \|f_m\|_{\mathcal{H}_m}^2} \leq \sum_{m=1}^M C_1 \lambda^{-\frac{s}{2}} \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}, \end{aligned}$$

where we used Young's inequality  $a^{1-s}b^s \leq (1-s)a + sb$  in the second line. Thus the RHS of the inequality (16) can be upper bounded by

$$\begin{aligned} &2C_1 \lambda^{-\frac{s}{2}} \mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \frac{\left| \frac{1}{n} \sum_{i=1}^n \sigma_i \left( \sum_{m=1}^M f_m(x_i) \right) \right|}{\sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right] \\ &\leq 2C_1 \lambda^{-\frac{s}{2}} \mathbb{E} \left[ \sup_{f_m \in \mathcal{H}_m} \max_m \frac{\left| \frac{1}{n} \sum_{i=1}^n \sigma_i f_m(x_i) \right|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}} \right], \end{aligned}$$

where we used the relation  $\frac{\sum_m a_m}{\sum_m b_m} \leq \max_m \left( \frac{a_m}{b_m} \right)$  for all  $a_m \geq 0$  and  $b_m \geq 0$  with a convention  $\frac{0}{0} = 0$ . Therefore, by  $\frac{\log(M)}{\sqrt{n}} \leq 1$  and Theorem 10 where  $\sigma_i$  is substituted into  $\epsilon_i$ , the right hand side is upper bounded by  $16KC_1(C_s + 1 + C_1)\lambda^{-\frac{s}{2}}\xi_n$ . Here we again apply Talagrand's concentration inequality, then we have

$$\begin{aligned} &P \left( \sup_{f_m \in \mathcal{H}_m} \frac{\left| \left\| \sum_{m=1}^M f_m \right\|_n^2 - \left\| \sum_{m=1}^M f_m \right\|_{L_2(\Pi)}^2 \right|}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \right. \\ &\quad \left. \geq K \left[ 16KC_1(C_s + 1 + C_1)\lambda^{-\frac{s}{2}}\xi_n + \sqrt{\frac{t}{n}} C_1 \lambda^{-\frac{s}{2}} + \frac{C_1^2 \lambda^{-s} t}{n} \right] \right) \leq e^{-t}, \end{aligned} \quad (17)$$

where we substituted the following upper bounds of  $B$  and  $U$ :

$$\begin{aligned} B^2 &= \sup_{f_m \in \mathcal{H}_m} \mathbb{E} \left[ \left( \frac{(\sum_{m=1}^M f_m)^2}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \right)^2 \right] \\ &\leq \sup_{f_m \in \mathcal{H}_m} \mathbb{E} \left[ \frac{(\sum_{m=1}^M f_m)^2}{\left( \sum_{m=1}^M \|f_m\|_{L_2(\Pi)} \right)^2} \frac{(\| \sum_{m=1}^M f_m \|_{\infty})^2}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \right] \\ &\leq \sup_{f_m \in \mathcal{H}_m} \frac{\left( \sum_{m=1}^M \|f_m\|_{L_2(\Pi)} \right)^2}{\left( \sum_{m=1}^M \|f_m\|_{L_2(\Pi)} \right)^2} \frac{(\sum_{m=1}^M C_1 \lambda^{-\frac{s}{2}} \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2})^2}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \\ &\leq C_1^2 \lambda^{-s}, \end{aligned}$$

where in the second inequality we used the relation  $\mathbb{E}[(\sum_{m=1}^M f_m)^2] = \mathbb{E}[\sum_{m,m'=1}^M f_m f_{m'}] \leq \sum_{m,m'=1}^M \|f_m\|_{L_2(\Pi)} \|f_{m'}\|_{L_2(\Pi)} = (\sum_{m=1}^M \|f_m\|_{L_2(\Pi)})^2$ , and

$$\begin{aligned} U &= \sup_{f_m \in \mathcal{H}_m} \left\| \frac{(\sum_{m=1}^M f_m)^2}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \right\| \leq \sup_{f_m \in \mathcal{H}_m} \frac{(\sum_{m=1}^M C_1 \lambda^{-\frac{s}{2}} \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2})^2}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \\ &\leq C_1^2 \lambda^{-s}. \end{aligned}$$

Now notice that

$$K \left[ 16KC_1(C_s + 1 + C_1)\lambda^{-\frac{s}{2}}\xi_n + \sqrt{\frac{t}{n}} C_1 \lambda^{-\frac{s}{2}} + \frac{C_1^2 \lambda^{-s} t}{n} \right]$$

$$\begin{aligned}
&\leq \sqrt{n}K \left[ 16KC_1(C_s + 1 + C_1) \frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \xi_n + \sqrt{\frac{t}{\log(M)}} C_1 \xi_n \sqrt{\frac{\log(M)}{n}} + \frac{C_1^2 \xi_n^2 t}{\sqrt{n}} \right] \\
&\leq \sqrt{n}K \left[ 16KC_1(C_s + 1 + C_1) + \sqrt{\frac{t}{\log(M)}} C_1 + \frac{C_1^2 t}{\sqrt{n}} \right] \xi_n^2.
\end{aligned}$$

Therefore Eq. (17) gives the following inequality

$$\begin{aligned}
&\sup_{f_m \in \mathcal{H}_m} \frac{\left| \left\| \sum_{m=1}^M f_m \right\|_n^2 - \left\| \sum_{m=1}^M f_m \right\|_{L_2(\Pi)}^2 \right|}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \\
&\leq K [16KC_1(C_s + 1 + C_1) + C_1 + C_1^2] \sqrt{n} \xi_n^2 \max(1, \sqrt{t/\log(M)}, t/\sqrt{n}).
\end{aligned}$$

with probability  $1 - \exp(-t)$ . By substituting  $\tilde{\phi}'_s = K [16KC_1(C_s + 1 + C_1) + C_1 + C_1^2]$  and  $t = \zeta_n(r, \lambda)$ , we have

$$\begin{aligned}
&\sup_{f_m \in \mathcal{H}_m} \frac{\left| \left\| \sum_{m=1}^M f_m \right\|_n^2 - \left\| \sum_{m=1}^M f_m \right\|_{L_2(\Pi)}^2 \right|}{\left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2} \\
&\leq \tilde{\phi}'_s \sqrt{n} \xi_n^2 \max \left( 1, \sqrt{\frac{\zeta_n(r, \lambda)}{\log(M)}}, \frac{\zeta_n(r, \lambda)}{\sqrt{n}} \right) \leq \tilde{\phi}'_s \sqrt{n} \xi_n^2 \max \left( 1, \frac{r}{\tilde{\phi}'_s \sqrt{n} \xi_n^2} \right) \leq \max \left( \tilde{\phi}'_s \sqrt{n} \xi_n^2, r \right).
\end{aligned}$$

with probability  $1 - \exp(-\zeta_n(r, \lambda))$ . ■

Now we define

$$\phi_s := \max(\tilde{\phi}'_s, \tilde{\phi}_s, 1) = \max(K [16KC_1(C_s + 1 + C_1) + C_1 + C_1^2], 2KL(C_s + 1 + C_1), 1),$$

where  $K$  is the universal constant appeared in Talagrand's concentration inequality (Proposition 5). We define events  $\mathcal{E}_1(t)$  and  $\mathcal{E}_2(r)$  as

$$\begin{aligned}
\mathcal{E}_1(t) &= \left\{ \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i) \right| \leq \eta(t) \phi_s \xi_n \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2}, \forall f_m \in \mathcal{H}_m, \forall m = 1, \dots, M \right\}, \\
\mathcal{E}_2(r) &= \left\{ \left| \left\| \sum_{m=1}^M f_m \right\|_n^2 - \left\| \sum_{m=1}^M f_m \right\|_{L_2(\Pi)}^2 \right| \leq \max(\phi_s \sqrt{n} \xi_n^2, r) \left( \sum_{m=1}^M \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2} \right)^2, \right. \\
&\quad \left. \forall f_m \in \mathcal{H}_m, \forall m = 1, \dots, M \right\}.
\end{aligned}$$

Theorems 10 and 11 give that  $P(\mathcal{E}_1(t)) \geq 1 - e^{-t}$  and  $P(\mathcal{E}_2(r)) \geq 1 - \exp(-\zeta_n(r, \lambda))$  under some conditions.

The next lemma gives a bound of irrelevant components ( $m \in I_0^c$ ) of  $\hat{f}$  in terms of the relevant components.

**Lemma 12** Set  $\lambda_1^{(n)} = 4\phi_s \eta(t) \xi_n(\lambda)$ ,  $\lambda_2^{(n)} = \lambda$ ,  $\lambda_3^{(n)} = \lambda$  for arbitrary  $\lambda > 0$ . Then for all  $n$  and  $r(\geq 0)$  such that  $\frac{\log(M)}{\sqrt{n}} \leq 1$  and  $\max(\phi_s \sqrt{n} \xi_n^2(\lambda), r) \leq \frac{1}{2}$ , we have

$$\begin{aligned}
&\sum_{m=1}^M \sqrt{\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} \\
&\leq 8 \sum_{m \in I_0} \left( 1 + \frac{\lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}} \right) \left( \sqrt{\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} \right), \tag{18}
\end{aligned}$$

with probability  $1 - \exp(-t) - \exp(-\zeta_n(r, \lambda))$ .

**Proof: (Lemma 12)** On the event  $\mathcal{E}_2(r)$ , for all  $f_m \in \mathcal{H}_m$  we obtain the upper bound of the regularization term as

$$\begin{aligned} & \sqrt{\|f_m\|_n^2 + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}^2} \\ & \leq \sqrt{\|f_m\|_{L_2(\Pi)}^2 + \max(\phi_s \sqrt{n} \xi_n^2(\lambda), r) (\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2) + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}^2} \\ & \leq \sqrt{\frac{3}{2} (\|f_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2)}, \end{aligned} \quad (19)$$

because  $\max(\phi_s \sqrt{n} \xi_n^2(\lambda), r) \leq \frac{1}{2}$  and  $\lambda = \lambda_2^{(n)}$ . On the other hand, we also obtain a lower bound as

$$\begin{aligned} & \sqrt{\|f_m\|_n^2 + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}^2} \\ & \geq \sqrt{\|f_m\|_{L_2(\Pi)}^2 - \max(\phi_s \sqrt{n} \xi_n^2(\lambda), r) (\|f_m\|_{L_2(\Pi)}^2 + \lambda \|f_m\|_{\mathcal{H}_m}^2) + \lambda_2^{(n)} \|f_m\|_{\mathcal{H}_m}^2} \\ & \geq \sqrt{\frac{1}{2} (\|f_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2)}, \end{aligned} \quad (20)$$

for all  $f_m \in \mathcal{H}_m$ .

Note that, since  $\hat{f}$  minimizes the objective function,

$$\begin{aligned} & \|\hat{f} - f^*\|_n^2 + \sum_{m=1}^M (\lambda_1^{(n)} \sqrt{\|\hat{f}_m\|_n^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2) \\ & \leq \frac{1}{n} \sum_{n=1}^n \sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) + \sum_{m \in I_0} (\lambda_1^{(n)} \sqrt{\|f_m^*\|_n^2 + \lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2). \end{aligned}$$

This implies

$$\begin{aligned} & \|\hat{f} - f^*\|_n^2 + \sum_{m \in I_0^c} \lambda_1^{(n)} \sqrt{\|\hat{f}_m\|_n^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2} \\ & \leq \frac{1}{n} \sum_{n=1}^n \sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) \\ & \quad + \sum_{m \in I_0} (\lambda_1^{(n)} \sqrt{\|f_m^* - \hat{f}_m\|_n^2 + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} (\|f_m^*\|_{\mathcal{H}_m}^2 - \|\hat{f}_m\|_{\mathcal{H}_m}^2)). \end{aligned}$$

Thus on the event  $\mathcal{E}_1(t)$  and  $\mathcal{E}_2(r)$ , by Eq. (19) and Eq. (20), we have

$$\begin{aligned} & \|\hat{f} - f^*\|_n^2 + \frac{1}{2} \sum_{m \in I_0^c} \lambda_1^{(n)} \sqrt{\|\hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2} \\ & \leq \sum_{m=1}^M \eta(t) \phi_s \xi_n \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} + \\ & \quad \sum_{m \in I_0} \left( \frac{3}{2} \lambda_1^{(n)} \sqrt{\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} (2 \langle f_m^*, f_m^* - \hat{f}_m \rangle_{\mathcal{H}_m} - \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2) \right) \\ & \Rightarrow \\ & \frac{1}{4} \sum_{m \in I_0^c} \lambda_1^{(n)} \sqrt{\|\hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2} \\ & \leq \sum_{m \in I_0} \left( \frac{7}{4} \lambda_1^{(n)} \sqrt{\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} + 2 \lambda_3^{(n)} \langle T_m^{\frac{q}{2}} g_m^*, f_m^* - \hat{f}_m \rangle_{\mathcal{H}_m} \right). \end{aligned}$$

Now by the Young's inequality for positive symmetric operator, we have

$$\lambda_3^{(n)1-q} T_m^q = \lambda_3^{(n)\frac{1}{2}} \left( \lambda_3^{(n)-\frac{1}{2}} T_m \lambda_3^{(n)-\frac{1}{2}} \right)^q \lambda_3^{(n)\frac{1}{2}}$$

$$\leq qT_m + (1-q)\lambda_3^{(n)}.$$

Thus

$$\begin{aligned}
& \lambda_3^{(n)} \langle f_m^*, f_m^* - \hat{f}_m \rangle_{\mathcal{H}_m} \\
&= \lambda_3^{(n)} \langle T_m^{\frac{q}{2}} g_m^*, f_m^* - \hat{f}_m \rangle_{\mathcal{H}_m} \\
&\leq \lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m} \|\lambda_3^{(n) \frac{1-q}{2}} T_m^{\frac{q}{2}} (f_m^* - \hat{f}_m)\|_{\mathcal{H}_m} \\
&\leq \lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m} \sqrt{\langle f_m^* - \hat{f}_m, (qT_m + (1-q)\lambda_3^{(n)}) f_m^* - \hat{f}_m \rangle} \\
&= \lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m} \sqrt{q\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + (1-q)\lambda_3^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} \\
&\leq \lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m} \sqrt{\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_3^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2}.
\end{aligned} \tag{22}$$

Therefore we have

$$\begin{aligned}
& \frac{1}{4} \sum_{m \in I_0^c} \lambda_1^{(n)} \sqrt{\|\hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2} \\
&\leq \sum_{m \in I_0} \left( \frac{7}{4} \lambda_1^{(n)} + 2\lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m} \right) \sqrt{\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2}.
\end{aligned}$$

with probability  $1 - \exp(-t) - \exp(-\zeta_n(r, \lambda))$ . The assertion is obvious from this bound.  $\blacksquare$

The next theorem immediately gives Theorem 2.

**Theorem 13** *Let  $\lambda_1^{(n)} = 4\phi_s\eta(t)\xi_n(\lambda)$ ,  $\lambda_2^{(n)} = \lambda$ ,  $\lambda_3^{(n)} = \lambda$  for arbitrary  $\lambda > 0$ . Then for all  $n$  and  $r(\geq 0)$  satisfying  $\frac{\log(M)}{\sqrt{n}} \leq 1$  and the following inequality:*

$$\frac{128 \max(\phi_s \sqrt{n} \xi_n^2(\lambda), r) \left( d + \frac{\lambda_3^{(n) 1+q}}{\lambda_1^{(n) 2}} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right)}{(1 - \rho(I_0)^2) \kappa(I_0)} \leq \frac{1}{8}, \tag{23}$$

we have

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{48}{(1 - \rho(I_0))^2 \kappa(I_0)} \left( d\lambda_1^{(n) 2} + \lambda_3^{(n) 1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right),$$

with probability  $1 - \exp(-t) - \exp(-\zeta_n(r, \lambda))$  for all  $t \geq 1$ .

**Proof: (Theorem 13)** By Eq. (21), we have

$$\begin{aligned}
& \|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m \in I_0^c} (\lambda_1^{(n)} \sqrt{\|\hat{f}_m\|_n^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2) + \sum_{m \in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \\
&\leq (\|\hat{f} - f^*\|_{L_2(\Pi)}^2 - \|\hat{f} - f^*\|_n^2) + \frac{1}{n} \sum_{n=1}^n \sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) \\
&\quad + \sum_{m \in I_0} (\lambda_1^{(n)} \sqrt{\|\hat{f}_m - f_m^*\|_n^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} 2\langle f_m^*, f_m^* - \hat{f}_m \rangle_m).
\end{aligned}$$

Here on the event  $\mathcal{E}_2(r)$ , the above inequality gives

$$\begin{aligned}
& \|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \frac{1}{2} \sum_{m \in I_0^c} (\lambda_1^{(n)} \sqrt{\|\hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2) + \sum_{m \in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \\
&\leq \max(\phi_s \sqrt{n} \xi_n^2, r) \left( \sum_{m=1}^M \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} \right)^2 + \frac{1}{n} \sum_{n=1}^n \sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) \\
&\quad + \sum_{m \in I_0} \left( \frac{3}{2} \lambda_1^{(n)} \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} + \lambda_3^{(n)} 2\langle f_m^*, f_m^* - \hat{f}_m \rangle_m \right).
\end{aligned} \tag{24}$$



Moreover notice that the assumption (23) implies  $\max(\phi_s \sqrt{n} \xi_n^2, r) \leq \frac{1}{2}$ . Thus Eq. (18) in Lemma 12 holds.

*Step 1.* (Bound of the first term in the RHS of Eq. (24)) By Eq. (18) in Lemma 12, the first term on the RHS of Eq. (24) can be upper bounded as

$$\begin{aligned}
& \max(\phi_s \sqrt{n} \xi_n^2, r) \left( \sum_{m=1}^M \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} \right)^2 \\
& \leq \max(\phi_s \sqrt{n} \xi_n^2, r) \left( 8 \sum_{m \in I_0} \left( 1 + \frac{\lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}} \right) \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} \right)^2 \\
& \leq 128 \max(\phi_s \sqrt{n} \xi_n^2, r) \left( d + \frac{\lambda_3^{(n) 1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{\lambda_1^{(n) 2}} \right) \sum_{m \in I_0} \left( \|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right) \\
& \leq 128 \max(\phi_s \sqrt{n} \xi_n^2, r) \left( d + \frac{\lambda_3^{(n) 1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{\lambda_1^{(n) 2}} \right) \left( \frac{\|\hat{f} - f^*\|_{L_2(\Pi)}^2}{(1 - \rho(I_0)^2) \kappa(I_0)} + \sum_{m \in I_0} \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right). \tag{25}
\end{aligned}$$

By assumption, we have  $128 \frac{\phi_s \max(\phi_s \sqrt{n} \xi_n^2, r)}{(1 - \rho(I_0)^2) \kappa(I_0)} \left( d + \frac{\lambda_3^{(n) 1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{\lambda_1^{(n) 2}} \right) \leq \frac{1}{8}$ . Hence the RHS of the above inequality is bounded by  $\frac{1}{8} \left( \|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m \in I_0} \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right)$ .

*Step 2.* (Bound of the second term in the RHS of Eq. (24)) By Eq. (18) in Lemma 12, we have on the event  $\mathcal{E}_1$

$$\begin{aligned}
& \frac{1}{n} \sum_{n=1}^n \sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) \leq \sum_{m=1}^M \eta(t) \phi_s \xi_n \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} \\
& \leq \sum_{m \in I_0} 8 \left( 1 + \frac{\lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}} \right) \eta(t) \phi_s \xi_n \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} \\
& \leq \frac{256 \phi_s^2 \eta(t)^2 \xi_n^2}{(1 - \rho(I_0)^2) \kappa(I_0)} \left( d + \frac{\lambda_3^{(n) 1+q}}{\lambda_1^{(n) 2}} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right) \\
& \quad + \frac{(1 - \rho(I_0)^2) \kappa(I_0)}{8} \sum_{m \in I_0} \left( \|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right) \\
& \leq \frac{16}{(1 - \rho(I_0)^2) \kappa(I_0)} \left( d \lambda_1^{(n) 2} + \lambda_3^{(n) 1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right) + \frac{1}{8} \left( \|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m \in I_0} \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right). \tag{26}
\end{aligned}$$

*Step 3.* (Bound of the third term in the RHS of Eq. (24)) By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
& \sum_{m \in I_0} \frac{3}{2} \lambda_1^{(n)} \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2} \\
& \leq \frac{9}{2(1 - \rho(I_0)^2) \kappa(I_0)} d \lambda_1^{(n) 2} + \frac{(1 - \rho(I_0)^2) \kappa(I_0)}{8} \sum_{m \in I_0} \left( \|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right) \\
& \leq \frac{9}{2(1 - \rho(I_0)^2) \kappa(I_0)} d \lambda_1^{(n) 2} + \frac{1}{8} \left( \|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m \in I_0} \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right). \tag{27}
\end{aligned}$$

*Step 4.* (Bound of the last term in the RHS of Eq. (24)) By Eq. (22), we have

$$\sum_{m \in I_0} 2 \lambda_3^{(n)} \langle f_m^*, f_m^* - \hat{f}_m \rangle_{\mathcal{H}_m} \leq 2 \sum_{m \in I_0} \lambda_3^{(n) \frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m} \sqrt{\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2}$$

$$\begin{aligned}
&\leq \frac{8 \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{(1 - \rho(I_0)^2)\kappa(I_0)} d\lambda_3^{(n)^{1+q}} + \frac{(1 - \rho(I_0)^2)\kappa(I_0)}{8} \sum_{m \in I_0} \left( \|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right) \\
&\leq \frac{8 \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{(1 - \rho(I_0)^2)\kappa(I_0)} d\lambda_3^{(n)^{1+q}} + \frac{1}{8} \left( \|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m \in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \right). \tag{28}
\end{aligned}$$

*Step 5.* (Combining all the bounds) Substituting the inequalities (25), (26), (27) and (28) to Eq. (24), we obtain

$$\begin{aligned}
&\frac{1}{2} \|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\
&\leq \frac{16}{(1 - \rho(I_0)^2)\kappa(I_0)} \left( d\lambda_1^{(n)^2} + \lambda_3^{(n)^{1+q}} R_{g^*}^2 \right) + \frac{9}{2(1 - \rho(I_0)^2)\kappa(I_0)} d\lambda_1^{(n)^2} + \frac{8R_{g^*}^2}{(1 - \rho(I_0)^2)\kappa(I_0)} \lambda_3^{(n)^{1+q}} \\
&\leq \frac{24}{(1 - \rho(I_0)^2)\kappa(I_0)} \left( d\lambda_1^{(n)^2} + \lambda_3^{(n)^{1+q}} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right).
\end{aligned}$$

This gives the assertion. ■

## E Proof of Theorem 3

**Proof: (Theorem 3)** The  $\delta$ -packing number  $\mathcal{M}(\delta, \mathcal{G}, L_2(P))$  of a function class  $\mathcal{G}$  with respect to  $L_2(P)$  norm is the largest number of functions  $\{f_1, \dots, f_{\mathcal{M}}\} \subseteq \mathcal{G}$  such that  $\|f_i - f_j\|_{L_2(P)} \geq \delta$  for all  $i \neq j$ . It is easily checked that

$$\mathcal{N}(\delta/2, \mathcal{G}, L_2(P)) \leq \mathcal{M}(\delta, \mathcal{G}, L_2(P)) \leq \mathcal{N}(\delta, \mathcal{G}, L_2(P)). \tag{29}$$

First we give the assertion about the  $\ell_\infty$ -mixed-norm ball (Eq. (10)). To simplify the notation, set  $R = R_\infty$ . For a given  $\delta_n > 0$  and  $\varepsilon_n > 0$ , let  $Q$  be the  $\delta_n$  packing number  $\mathcal{M}(\delta_n, \mathcal{H}_{\ell_\infty}^{d,q}(R), L_2(\Pi))$  of  $\mathcal{H}_{\ell_\infty}^{d,q}(R)$  and  $N$  be the  $\varepsilon_n$  covering number  $\mathcal{N}(\varepsilon_n, \mathcal{H}_{\ell_\infty}^{d,q}(R), L_2(\Pi))$  of  $\mathcal{H}_{\ell_\infty}^{d,q}(R)$ . Raskutti et al. (2010) utilized the techniques developed by Yang and Barron (1999) to show the following inequality in their proof of Theorem 2(b) :

$$\begin{aligned}
\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R)} \mathbb{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] &\geq \inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R)} \frac{\delta_n^2}{2} P[\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \geq \delta_n^2/2] \\
&\geq \frac{\delta_n^2}{2} \left( 1 - \frac{\log(N) + \frac{n}{2\sigma^2} \varepsilon_n^2 + \log(2)}{\log(Q)} \right).
\end{aligned}$$

Now let  $\tilde{Q}_m := \mathcal{M}(\delta_n/\sqrt{d}, \mathcal{H}_m^q(R), L_2(\Pi))$  (remind the definition of  $\mathcal{H}_m^q(R)$  (Eq. (11)), and since now  $\mathcal{H}_m$  is taken as  $\tilde{\mathcal{H}}$  for all  $m$ , the value  $\tilde{Q}_m$  is common for all  $m$ ). Thus by taking  $\delta_n$  and  $\varepsilon_n$  to satisfy

$$\frac{n}{2\sigma^2} \varepsilon_n^2 \leq \log(N), \tag{30}$$

$$4 \log(N) \leq \log(Q), \tag{31}$$

the minimax rate is lower bounded by  $\frac{\delta_n^2}{4}$ . In Lemma 5 of Raskutti et al. (2010), it is shown that if  $\tilde{Q}_1 \geq 2$  and  $d \leq M/4$ , we have

$$\log(Q) \sim d \log(\tilde{Q}_1) + d \log\left(\frac{M}{d}\right).$$

By the estimation of the covering number of  $\mathcal{H}_m^q(1)$  (Eq. (12)), the strong spectrum assumption (Eq. (8)) and the relation (29), we have

$$\log(\tilde{Q}_1) \sim \left( \frac{\delta_n}{R\sqrt{d}} \right)^{-2\frac{s}{1+q}} = \left( \frac{\delta_n}{R\sqrt{d}} \right)^{-2\tilde{s}}.$$

Thus the conditions (31) and (30) are satisfied if we set  $\delta_n = C\varepsilon_n$  with an appropriately chosen constant  $C$  and we take  $\varepsilon_n$  so that the following inequality holds:

$$n\varepsilon_n^2 \lesssim d^{1+\tilde{s}} R^{2\tilde{s}} \varepsilon_n^{-2\tilde{s}} + d \log\left(\frac{M}{d}\right).$$

It suffices to take

$$\varepsilon_n^2 \sim dn^{-\frac{1}{1+s}} R^{\frac{2s}{1+s}} + \frac{d \log \left( \frac{M}{d} \right)}{n}. \quad (32)$$

Note that we have taken  $R \geq \sqrt{\frac{\log(M/d)}{n}}$ , thus  $\tilde{Q}_m \geq 2$  is satisfied if we take the constant in Eq. (32) appropriately. Thus we obtain the assertion (10).

Next we give the assertion about the  $\ell_2$ -mixed-norm ball (Eq. (9)). To simplify the notation, set  $R = R_2$ . Since  $\mathcal{H}_{\ell_2}^{d,q}(R) \supseteq \mathcal{H}_{\ell_\infty}^{d,q}(R/\sqrt{d})$ , we obtain

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_2}^{d,q}(R)} \mathbb{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] \geq \inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R/\sqrt{d})} \mathbb{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2].$$

Here notice that we have  $\frac{R}{\sqrt{d}} \geq \sqrt{\frac{\log(M/d)}{n}}$  by assumption. Thus we can apply the assertion about the  $\ell_\infty$ -mixed-norm ball (10) to bound the RHS of the just above display. We have shown that

$$\begin{aligned} \inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R/\sqrt{d})} \mathbb{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] &\gtrsim dn^{-\frac{1}{1+s}} (R/\sqrt{d})^{\frac{2s}{1+s}} + \frac{d \log \left( \frac{M}{d} \right)}{n} \\ &= d^{\frac{1}{1+s}} n^{-\frac{1}{1+s}} R^{\frac{2s}{1+s}} + \frac{d \log \left( \frac{M}{d} \right)}{n}. \end{aligned}$$

This gives the assertion (9). ■

## References

- A. Argyriou, R. Hauser, C. A. Micchelli, and M. Pontil. A dc-programming algorithm for kernel selection. In *the 23rd International Conference on Machine Learning*, 2006.
- F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.
- F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 105–112. 2009.
- F. R. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *the 21st International Conference on Machine Learning*, pages 41–48, 2004.
- C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.
- O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. *C. R. Acad. Sci. Paris Ser. I Math.*, 334:495–500, 2002.
- E. Candes and T. Tao. The Dantzig selector: Statistical estimation when  $p$  is much larger. *The Annals of Statistics*, 35(6):2313–2351, 2007.
- A. Caponnetto and E. de Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1/3):131, 2002.
- C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 396–404. 2009a.
- C. Cortes, M. Mohri, and A. Rostamizadeh.  $L_2$  regularization for learning kernels. In *the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, 2009b. Montréal, Canada.
- G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

- M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate  $\ell_p$ -norm multiple kernel learning. In *Advances in Neural Information Processing Systems 22*, pages 997–1005, Cambridge, MA, 2009. MIT Press.
- M. Kloft, U. Rückert, and P. L. Bartlett. A unifying view of multiple kernel learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2010.
- V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of the Annual Conference on Learning Theory*, pages 229–238, 2008.
- V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6): 3660–3695, 2010.
- G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.
- M. Ledoux and M. Talagrand. *Probability in Banach Spaces. Isoperimetry and Processes*. Springer, New York, 1991. MR1102015.
- L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol*, 70:53–71, 2008.
- L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.
- C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.
- C. S. Ong, A. J. Smola, and R. C. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071, 2005.
- G. Raskutti, M. Wainwright, and B. Yu. Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. In *Advances in Neural Information Processing Systems 22*, pages 1563–1570. MIT Press, Cambridge, MA, 2009.
- G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Technical report, 2010. arXiv:1008.3654.
- B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- J. Shawe-Taylor. Kernel learning for novelty detection. In *NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, Whistler, 2008.
- J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Proceedings of the Annual Conference on Learning Theory*, 2006.
- I. Steinwart. *Support Vector Machines*. Springer, 2008.
- I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pages 79–93, 2009.
- M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126: 505–563, 1996.
- R. Tomioka and T. Suzuki. Sparsity-accuracy trade-off in MKL. In *NIPS 2009 Workshop: Understanding Multiple Kernel Learning Methods*, Whistler, 2009.
- A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.
- M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *The 26th International Conference on Machine Learning*, 2009.

- Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.
- Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In S. Dasgupta and A. Klivans, editors, *Proceedings of the Annual Conference on Learning Theory*, Montreal Quebec, 2009. Omnipress.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical: Series B*, 67(2):301–320, 2005.